Directional Self-supervised Learning for Heavy Image Augmentations

Yalong Bai[1*] Yifan Yang[2*] Wei Zhang[1†] Tao Mei[1]
[1]JD Explore Academy [2]Peking University
ylbai@outlook.com, yif_yang@pku.edu.cn, wzhang.cu@gmail.com, tmei@jd.com

Abstract

Despite the large augmentation family, only a few cherry-picked robust augmentation policies are beneficial to self-supervised image representation learning. In this paper, we propose a directional self-supervised learning paradigm (DSSL), which is compatible with significantly more augmentations. Specifically, we adapt heavy augmentation policies after the views lightly augmented by standard augmentations, to generate harder view (HV). HV usually has a higher deviation from the original image than the lightly augmented standard view (SV). Unlike previous methods equally pairing all augmented views to symmetrically maximize their similarities, DSSL treats augmented views of the same instance as a partially ordered set (with directions as SV↔SV, SV←HV), and then equips a directional objective function respecting to the derived relationships among views. DSSL can be easily implemented with a few lines of codes and is highly flexible to popular self-supervised learning frameworks, including SimCLR, SimSiam, BYOL. Extensive experimental results on CIFAR and ImageNet demonstrated that DSSL can stably improve various baselines with compatibility to a wider range of augmentations. Code is available at: https://github.com/Yif-Yang/DSSL.

1. Introduction

Unsupervised visual representation learning aims at learning image features without using manual semantic annotations. Recently, self-supervised learning driven by instance discrimination tasks [3, 13, 22, 27, 29] or Siamese architecture [5, 12] has achieved great success in learning high-quality visual features and closing the performance gap with supervised pretraining on various computer tasks. The visual embedding space of self-supervised learning method is constructed by minimizing the dissimilarity among representations of variations derived from the same image, and/or increasing the distance between the representations of augmented view from different images (negative pair). Image transformation, which aims to generate variations from the same image [3, 5, 12, 13], plays a crucial role in the self-supervised visual representations learning.

Augment	SimSiam	w/ DSSL
Standard Aug.	92.17	
w/ *JigSaw*(2)	92.79	**93.56**
w/ *JigSaw*(4)	89.72	**91.95**
w/ *RA*(1,1)	88.00	**92.29**
w/ *RA*(2,1)	82.80	**93.09**
w/ *RA*(2,5)	9.78	**94.17**
w/ *UA*	91.11	**93.27**

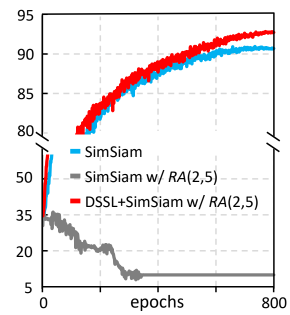

Figure 1. Left: Linear evaluation accuracy of SimSiam [5] on CIFAR-10 by adding extra *heavy augmentations* besides the original *standard augmentations*. The number of grids for *JigSaw*(n) is $n \times n$. *RA*(m, n) is the RandAugment [8] with n augmentation transformation of m magnitude. *UA* denotes the UniformAumgent [19]. Right: validation accuracy of kNN classification during pre-training. Incorporating heavy augmentations on SimSiam results in unstable performance even *collapsing* (linear evaluation on collapsed model tends to random guess). Our DSSL consistently benefits from heavy augmentations for higher performances.

However, these self-supervised learning methods share a common fundamental weakness: only a few carefully selected augmentation policies with proper settings are beneficial to the model training. The combination of *random crop, color distortion, Gaussian blur and grayscale* is crucial to achieving good performance, applied as the basic augmentation setting for many popular instance-wise self-supervised learning methods [2, 5, 12, 13, 17]. Here we define these augmentations as the **standard augmentations**. We denote the images augmented by standard augmentations as the standard views. Recent works find that a few other augmentations (e.g., *RandAugment*, *JigSaw*) can further improve the performance of self-supervised learning methods rely on negative pairs [23]. However, for negative pair free self-supervised learning methods like SimSiam [5], introducing such augmentations usually results in

*Equal contribution.
†Corresponding author.

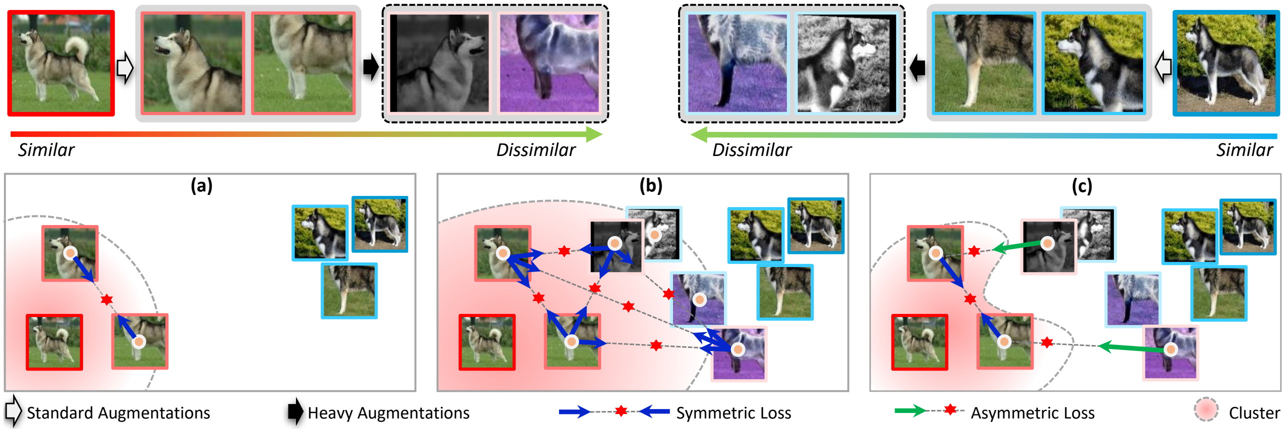

Figure 2. Overview of standard self-supervised learning and our DSSL. Original standard image transformations generate the standard views, and the harder view is derived from the the standard view by applying heavy augmentation *RandAugment*. (a) Standard instance-wise learning with standard views. (b) Instance-wise self-supervised learning after introducing heavily augmented (harder) views. Applying symmetric loss to maximize the similarity between standard and heavily augmented views roughly expands the feature cluster in the visual embedding space. The model may confuse the instance-level identity. (c) DSSL: To prevent the adverse effect from *missing information* of heavily augmented views, DSSL avoids arbitrarily maximizing their visual agreement. To tighten the feature cluster, DSSL applies an asymmetric loss for only gathering each heavily augmented view to its relevant standard view.

a lousy performance, even model collapsing during training, as shown in Fig. 1. We name these unstable and risky data augmentation policies as **heavy augmentations**, since they usually largely alters the image appearance.

Inspired by previous works [5, 17] that formulating the instance-wise self-supervised learning as K-means clustering of all augmented views from the same instance, we hypothesize a gold standard feature cluster for all views of one given image instance existing in the visual feature embedding space, and define d as the deviation of augmented view's feature from the core-point of its relevant gold standard feature cluster. Standard self-supervised learning methods treat all augmented views of the same image fairly to construct training pairs. As shown in Fig. 2 (a), such a strategy works well, and the model can converge stably for the standard image transformations. However, after incorporating the views augmented from heavy image transformations (Fig. 2 (b)), two obvious risks arise. 1) Closing the representation of standard views to heavily augmented views would roughly expand the feature cluster in the embedding space. This would increase the difficulty of constructing an embedding space where all instances are well-separated and also may cause unexpected confusion with other instance distributions [1, 26]. 2) Maximizing the visual agreement among heavily augmented views disaccords with the "InfoMin principle" [23]. Since the mutual information among views with large d is usually low, contrasting these views leads to *missing information* and results in poor performance in downstream tasks.

To address this, we propose Directional Self-supervised Learning (DSSL), a new training method for unsupervised representation learning that could stably improve the performance of instance-wise self-supervised learning by completely applying more heavy image transformations. Fig. 2 (c) shows an illustration of DSSL. For each standard view (SV) augmented from the original robust image transformations, we can generate various harder views (HV) derived from it by applying additional heavy augmentation policies. These heavily augmented view has a larger d than its relevant standard view. In this way, we can treat all augmented views of the same image as a partially ordered set (SV$\leftrightarrow$SV, SV$\leftarrow$HV) in terms of d. An asymmetric loss is introduced to encourage the representation of each heavily augmented view (HV) to be close to its relevant source standard view (SV). In this way, the feature cluster for all augmented views can be presented as non-convex, rather than the K-means convex clustering, the whole cluster is tightened. Moreover, DSSL discards the instance-wise self-supervised learning among RVs to bypass the issue of low mutual information among HVs. As a result, more augmentation policies can be introduced to enrich the information of the whole embedding space but keep the instances still well-separated.

DSSL is a straightforward algorithm that can be easily implemented in a few line of Pseudo code. Also, there are no additional hyper-parameter needs to be adjusted in DSSL. We validate the effectiveness of DSSL by evaluating it on several self-supervised benchmarks. In particular, on the ImageNet linear evaluation protocol, DSSL achieved stable performance improvements. All DSSL based pre-training models surpass the supervised pre-training model on the CIFAR-10 linear evaluation. Moreover, the transfer performance on detection and segmentation task further demonstrate the efficiency of DSSL with heavy augmentations.

The main contributions are summarized as follows:

- A novel Directional Self-supervised Learning (DSSL) paradigm is proposed for unsupervised visual representation learning. We introduce a partially ordered set to organize the augmented views, and introduce an asymmetric loss for harnessing rich information from heavy augmented views.
- DSSL is easy-to-implement and applicable to various standard instance-wise self-supervised learning frameworks by introducing minor modifications without any hyper-parameters.
- DSSL stably improves over various self-supervised learning methods on standard benchmarks, even when thoroughly applying heavy image transformations that show adverse effect to previous methods.

2. Related Work

Our work is related to studies in the instance-wise self-supervised learning and data augmentation policies.

Instance-wise self-supervised learning. Instance-level classification task treats each image and its variants as one specific class. It aims to construct visual embedding space by pull all samples in the same class close while pushing samples from other classes away. Since it is hard to directly categorize all training samples into a large number of classes [10], the early instance-wise contrastive learning method replaces the classifier with a memory bank [27] to store the previous features of all samples calculated in the previous stage and sample positive and negative pairs from the memory bank. Several other technologies have also been adopted and extended based on this method, such as introducing local similarity [32] and neighborhood discovery [15] for further improving the quality of feature embedding. He *et al*. [13] enhance the training of memory bank based contrastive learning model by storing representations from a momentum encoder instead of the trained network. Instead of storing previously computed representations, some other methods explore different instances' features within the current batch for negative sampling, and it requires a large batch size to work well [3, 22, 29].

All of the above methods require either a large batch, memory bank, or queue to provide enough negative samples for clustering or discriminating. More recently, some works have proposed advancing self-supervised pretraining without using negative samples, e.g., BYOL [12], SimSiam [5]. These negative pair free self-supervised learning methods are more resilient to the changes in the batch size and more friendly to low-resource implementations.

Data augmentation policies. The composition of multiple data augmentation operations is crucial in defining the contrastive prediction tasks that yield effective representations [3, 5, 12]. Until now, most of the high-performance contrastive learning framework are designed to learn representations by maximizing agreement between differently augmented views of the same image via a contrastive loss in the feature embedding space. However, different with the supervised learning methods which can benefit from various complex data augmentation polices [6–9, 30], there are only a few light augmentation policies playing as key contributors of the good performance of instance-wise self-supervised learning [2,3,5,12]. The composition of the random crop, optional left-right flip, color distortion, Gaussian blur is treated as a standard and robust augmentation setting for generating augmented views of training images in unsupervised visual representation learning methods [2, 5]. Also, the

Our experimental study also shows that directly applying complex/heavy data augmentation policies leads to damaging performance drop or even model collapsing for negative pair free instance-wise self-supervised learning methods. These heavy data augmentations construct views with small mutual information among them. According to the infoMin principle [23, 24], unsupervised learning methods trained on such views would result in the "missing information" regime of performance. Even such heavily augmented views have been demonstrated containing rich information [25] but they still may mislead the feature clustering in the embedding space.

Different with another concept of directional self-supervision loss proposed in [28] for exploiting the output consistency across different resolutions in 3D human pose estimation task, in this paper, we propose a general self-supervised learning framework DSSL for introducing various image transformations for instance-wise self-supervised learning with better theoretical justification. The contrast among heavily augmented views with strong probabilistic of missing information is disabled. Moreover, DSSL regards instance-wise self-supervised learning as optimizing a non-convex clustering task. An asymmetric loss is proposed for tightening feature clusters. As a result, DSSL can achieve stable performance improvements on various instance-wise self-supervised learning methods owing to the rich information from heavy image transformations and the data characteristic-based learning strategies.

3. Method

Instance-wise self-supervised learning methods aim to learn representation by maximizing agreement among differently augmented views of the same data example in the latent visual feature space. To ease the discussion, we start by briefly summarizing the standard instance-wise self-supervised learning with a unified formulation.

3.1. A unified formulation

Following the basic settings of recent works, the standard instance-wise self-supervised learning framework has four main components:

- A *data augmentation module* consisting with augmentation policies set $\mathcal{T}$ for generating augmented view for given image.
- Deep neural network *encoder* $f(\cdot)$ for projecting the input images to the latent space.
- *Projection head* $g(\cdot)$ for mapping the outputs of encoder networks to space where instance-wise self-supervised loss is applied.
- A *self-supervised loss* function defined for an instance-wise discrimination task or a feature prediction task.

Given an input image I without annotation, the data augmentation module produces augmented view pair set as

$$\mathbf{V}_{\mathcal{T}} = \{(t(I), t'(I)) \mid t, t' \sim \mathcal{T}\} . \tag{1}$$

where t and t' are random augmentation sampled from $\mathcal{T}$. During training, augmented view pair (v, v') is sampled from $\mathbf{V}_{\mathcal{T}}$. One of the augmented views v is feed into the encoders to get their visual representations $f(v)$. The projection head transforms the feature of augmented view into a vector as $z \triangleq g(f(v))$. The objective functions of instance-wise self-supervised learning is maximizing the agreement between augmented view pairs in $\mathbf{V}_{\mathcal{T}}$: $\mathcal{S}(z, y(v'))$, where y works as an operation to generate the label for neural network training.

For different instance-wise self-supervised learning methods, the function of $\mathcal{S}$ and operation y are implemented in different ways:

- In SimCLR [3], $y(v') = g(f(v'))$, and $\mathcal{S}$ is formulated softmax-style function across v, v' and the negative views v^- augmented from other images in the minbatch.
- In BYOL [12], $y(v') = f_\xi(v')$, where $f_\xi(\cdot)$ is a siamese encoder with the exponential moving average weights of f. $\mathcal{S}$ measures the mean squared error between the normalized z and $y(v')$.
- In SimSiam [5], $y(v') = f(v')$, but there is no gradient backward through the neural pathway computing $y(v')$. $\mathcal{S}$ measures the negative cosine similarity of given feature vector pair.

Since the augmented view is generated randomly, above objective function is implemented as **symmetrical** form:

$$\mathcal{L}_S = \mathcal{S}(z, y(v')) + \mathcal{S}(z', y(v)), \tag{2}$$

where $z' \triangleq g(f(v'))$.

3.2. Partially-ordered views construction

In this paper, we introduce additional heavy data augmentation policies $\widehat{\mathcal{T}}$, besides the previous carefully selected standard data augmentation polices $\mathcal{T}$. We advance the data augmentation module for generating augmented views from $\widehat{\mathcal{T}}$ and $\mathcal{T}$ jointly. In particular, a new augmented view $\widehat{v} \triangleq \widehat{t}(v)$ can be produced based on $v \triangleq t(I)$ by applying image augmentation $\widehat{t} \sim \widehat{\mathcal{T}}$, $t \sim \mathcal{T}$. It means each light augmented view v would have various relevant harder augmented view $\widehat{v}$, and $\widehat{v}$ is derived from v.

For a well-trained unsupervised learning model, all views of the same image should be clustered closely in the visual embedding space [2, 5, 17]. We define the deviation of augmented view's feature $f(v)$ from the core-point (original view I) of its relevant cluster as $d(v)$. It is not easy to measure the deviation degree quantitatively. But, in general, larger distortion magnitudes of generating v leads to larger $d(v)$. Applying additional heavy augmentation policies on a standard augmented view can generate a new view with a higher deviation from the original view. As there is no policy overlap between $\mathcal{T}$ and $\widehat{\mathcal{T}}$, and the operation of $\widehat{t}$ is non-identity. Thus, we can compare the relative magnitudes of d between two views that $d(\widehat{v}) > d(v)$, since $\widehat{v}$ is produced by applying additional different heavy augmentation operations $\widehat{t}$ on v, as shown in Fig. 3 (Left).

In this way, we can construct a directed training pair collection $\mathbf{V}_{\mathcal{T} \leftarrow \widehat{\mathcal{T}}}$ from a partially ordered augmented view set as

$$\mathbf{V}_{\mathcal{T} \leftarrow \widehat{\mathcal{T}}} = \left\{ \left(t(I), \widehat{t}(t(I))\right) \mid t \sim \mathcal{T}, \widehat{t} \sim \widehat{\mathcal{T}} \right\} . \tag{3}$$

After combining the original training pairs generated from $\mathcal{T}$, we get the final augmented view pairs collection for training as $\mathbf{V}_{\mathcal{T}} \cup \mathbf{V}_{\mathcal{T} \leftarrow \widehat{\mathcal{T}}}$. In particular, $\mathbf{V}_{\mathcal{T}}$ can be regarded as the collection of edges in a complete undirected graph whose vertices are the augmented views. While $\mathbf{V}_{\mathcal{T} \leftarrow \widehat{\mathcal{T}}}$ can be regarded as a collection of the directed edges in a directed acyclic graph, where the relations among vertices are measured by the magnitude of $d(\cdot)$.

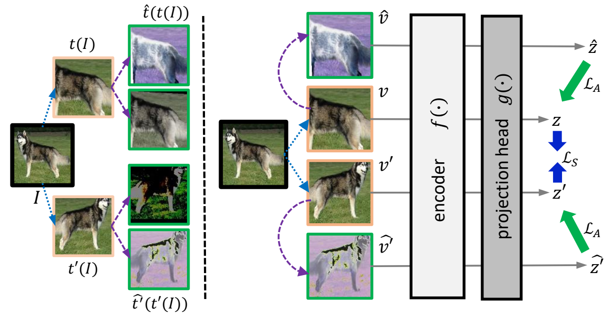

Figure 3. Illustration of our directional self-supervised learning (DSSL) framework. Left: Construction of partially ordered views. Right: Symmetric loss $\mathcal{L}_S$ for bidirectionally maximizing the agreement among augmented view pairs sampled from $\mathbf{V}_{\mathcal{T}}$ remains same. Asymmetric loss $\mathcal{L}_A$ is proposed for encouraging the representations of the heavily augmented views to be close to their source standard views, respecting the partially ordered relationship in $\mathbf{V}_{\mathcal{T} \to \widehat{\mathcal{T}}}$.

Algorithm 1 DSSL for SimSiam Pseudocode, PyTorch-like

```
# f: feature encoder; g: prediction head;
for I in dataloader:
    v, v′ = t(I), t′(I)  # standard data augmentation
    v̂, v̂′ = t̂(I), t̂′(I)  # heavy data augmentation
    y, y′, ŷ, ŷ′ = f(v), f(v′), f(v̂), f(v̂′)
    z, z′, ẑ, ẑ′ = g(y), g(y′), g(ŷ), g(ŷ′)
    # loss computation
    L = D(z, y′)/4 + D(z′, y)/4 + D(ẑ, y)/4 + D(ẑ′, y′)/4

    L.backward()  # back-propagate
    update(f, h)  # SGD update

def D(z, y):  # negative cosine similarity
    y = y.detach()  # stop gradient
    # l2-normalize
    z = normalize(z, dim=1)
    y = normalize(y, dim=1)
    return -(z * y).sum(dim=1).mean()
```

3.3. Directional self-supervised learning

For every two views v and v' sampled from $\mathbf{V}_{\mathcal{T}}$, we sample two directed augmented view pairs $(v, \widehat{v})$, $(v', \widehat{v'})$ from $\mathbf{V}_{\mathcal{T}\leftarrow\widehat{\mathcal{T}}}$. For the partially ordered relationship between $\widehat{v}$ and v, we introduce a $v \leftarrow \widehat{v}$ directional asymmetric loss as

$$\mathcal{L}_A = \mathcal{D}(\widehat{z}, y(v)) + \mathcal{D}(\widehat{z'}, y(v')), \tag{4}$$

where $\widehat{z} \triangleq g(f(\widehat{v}))$ and $\widehat{z'} \triangleq g(f(\widehat{v'}))$. The operation of $y(\cdot)$ stays the same with Eq. (2) following the settings in standard instance-wise self-supervised learning methods, but $y(\cdot)$ is only computed over the standard views in $\mathcal{L}_A$. This makes the self-supervised learning to be directional and asymmetrical. It means the optimization objective of $\mathcal{L}_A$ is to force the representation of heavily augmented view close to the representation of its relevant source view. $\mathcal{D}$ simply measure the negative cosine similarity among two vectors:

$$\mathcal{D}(z, y) = -\frac{\langle z, y\rangle}{\|z\|_2 \, \|y\|_2}.$$

As shown in Fig. 3 (Right), the objective function of our proposed directional self-supervised learning for given two standard views and their relevant harder views can be expressed as:

$$\mathcal{L}_{DSSL} = \mathcal{L}_S + \mathcal{L}_A. \tag{5}$$

The total loss of DSSL can be measured by $\sum_{\mathbf{V}_{\mathcal{T}}} \mathcal{L}_S + \sum_{\mathbf{V}_{\mathcal{T}\leftarrow\widehat{\mathcal{T}}}} \mathcal{L}_A$ averaged over all images. DSSL can be easily implemented on various instance-wise self-supervised learning frameworks in only a few lines of Pseudocode. Algorithm 1 shows the Pseudocode applying DSSL for unsupervised learning framework SimSiam [5].

Relation to previous self-supervised learning methods. We formulate the general self-supervised learning that considering all combinations of views augmented by $\mathcal{T}$ and $\widehat{\mathcal{T}}$ jointly as:

$$\alpha \sum_{\mathbf{V}_{\mathcal{T}}} \mathcal{L}_S + \beta \sum_{\mathbf{V}_{\widehat{\mathcal{T}}}} \mathcal{L}_S + \gamma \sum_{\mathbf{V}_{\mathcal{T}\leftarrow\widehat{\mathcal{T}}}} \mathcal{L}_A + \delta \sum_{\mathbf{V}_{\widehat{\mathcal{T}}\leftarrow\mathcal{T}}} \mathcal{L}_A, \tag{6}$$

where $\mathbf{V}_{\widehat{\mathcal{T}}}$ and $\mathbf{V}_{\widehat{\mathcal{T}}\leftarrow\mathcal{T}}$ are the augmented view collections constructed according to Eq. (1) and Eq. (3) respectively. $\leftarrow$ indicates the direction of predicting or contrasting. The α, β, γ and δ are the weights for each part of the loss. Previous self-supervised learning methods fairly treat all augmented views, thus the objective functions of these methods can be regards as the situation that $\alpha = \gamma = \beta = \delta = 1$. While the loss function of DSSL ($\mathcal{L}_{DSSL}$) can be regarded as the situation with $\alpha = \gamma = 1, \beta = \delta = 0$.

Considering that the view augmented from additional heavy augmentation policies $\widehat{\mathcal{T}}$ would tend to lose more information than its source view augmented from $\mathcal{T}$, a natural solution is to fit a lower confidence score for heavily augmented views when they play as the targets for optimizing the self-supervised loss. Thus, an ideal settings of the loss weights should ensure $\alpha, \gamma > \beta, \delta$. Further, we designed various self-supervised learning paradigms in Sec. 4.3 to analyze the positive/negative impact of each components in Eq. (6) and demonstrate the necessity of asymmetric loss for heavy augmentations in Eq. (5).

4. Experimental Results

4.1. Implementation details

Standard augmentations. We use the same sequence of image augmentations as in previous instance-wise self-supervised learning methods [3, 5, 12], including random *cropping and resizing*, *horizontal flip*, *color distortion*, *grayscale* and *Gaussian blur*. Each of the above augmentations has been proved effective in at least one typical instance-wise self-supervised learning method. The composition of all above image augmentations are treated as the standard augmentation $\mathcal{T}$.

Heavy augmentations. Inspired by the settings in InfoMin [23], we use *RandAugment* [8] and *Jigsaw* [6, 20] as heavy augmentations $\widehat{\mathcal{T}}$. These two augmentations have proven effective for supervised representation learning, and negative pairs required instance-wise contrastive learning. But we find that these augmentations result in lousy performance or even model collapsing for the negative-pairs-free unsupervised learning. We denote *RandAugment* as *RA*(n, m), where n is the number of augmentation policies randomly selected from 14 predefined augmentations, m is the magnitude for all the transformation. Unless otherwise specified, we equip a sequential combination of *RA*(2,5) and *Jigsaw* with 4×4 grids as heavy augmentation for experimental results reported in this paper.

Compared methods. We compare three typical self-supervised learning frameworks, including SimSiam, BYOL, and SimCLR, as shown in Table 1. SimCLR uses the normalized temperature-scaled cross entropy as $\mathcal{S}$, and the comparisons among positive and negative views are required for model training. Both BYOL and SimSiam are

Method	negative pairs	stop gradient	momentum encoder	reported	repro.
SimCLR [3]	✓			60.1	59.5
SimSiam [5]		✓		67.7	67.4
BYOL [12]		✓	✓	66.5	67.8

Table 1. Three instance-wise self-supervised learning methods used for comparisons and analysis in this paper.

negative sample free methods, and they stop the gradient of the neural branch for computing the label y, except BYOL that applies a momentum encoder for updating y.

For a fair comparison, we generated two augmented view pairs per image each time for negative sample free methods (SimSiam and BYOL) to keep the number of feature prediction pairs the same with their corresponding *w/ DSSL* versions. Moreover, for SimCLR that considers negative samples in the whole min-batch (batch size of n), its *w/ DSSL* version only adds one additional heavily augmented view with asymmetric loss item for each sample. The computation slightly increases from $2n^2$ to $2n^2+2n$ for each min-batch after applying DSSL on SimCLR. The number of feedforward and backward passes remains the same. More implementation details of compared methods and DSSL can be found in the Appendix.

Training details. Following the practice of the previous self-supervised learning methods, we use ResNet-50 and ResNet-18 as the basic feature encoder f for the experiments on ImageNet ILSVRC-2012 [21] and CIFAR-10 [16] dataset, respectively. We strictly follow the network architecture of projection head g, initialization, optimizer represented in these methods' original paper. Except that we apply the same SGD optimizer and learning rate schedule of SimSiam to BYOL since it can slightly improve the performance of BYOL. Existing works differ considerably in batch size and training epochs, which could significantly influence performance. We, therefore, compare all models of the same batch size of 512 and training epochs of 100 for unsupervised training on ImageNet, which is a resource-friendly implementation.

We elaborate the implementation details of the compared methods below.

- **SimCLR**: We used the PyTorch repo[1] officially recommended by the authors.
- **BYOL**: Our re-implemented BYOL has higher linear evaluation accuracy than the BYOL (100ep) results of ImageNet linear evaluation reported in the SimSiam paper (67.8% vs. 66.5%).
- **SimSiam**: To align settings of all compared methods in this paper, we used *lr* with cosine decay but did not fix the *lr* of prediction MLP during linear evaluation. The result of our reproduction is 67.4% on ImageNet (vs. 67.7% reported in the original paper). This performance gap is acceptable.

[1] https://github.com/AndrewAtanov/simclr-pytorch

4.2. Main results and discussions

Linear evaluation on ImageNet. We apply linear evaluation to measure the quality of the representations of DSSL based models after self-supervised pretraining on the unlabelled training images of the ImageNet dataset. Specifically, we train a linear classifier on top of the pre-trained representation. During training, the parameters of the backbone network (feature encoder) are frozen, while only the last fully connected layer is updated via backpropagation.

Tab. 2 reports the top-1 accuracy of compared methods and their DSSL versions. Maximizing the dissimilarity among different instances is a direct and effective way to learn a well-separated visual embedding space. Such a manner, up to a point, is robust to the missing-information hard views. However, for the negative pairs free method SimSiam, the heavy augmentations usually result in the model collapsing. Momentum encoder can increase model training stability since momentum would neutralize the misleading information from inconsistent representations. BYOL is more robust to the hard views than SimSiam. Although the improvement of DSSL on BYOL and SimCLR is limited under current augmentation setting, such limitation is easy to break after introducing more and heavier augmentations, as shown in Tab. 3 and Tab. 4. We further analyse the robustness of these mechanisms in the later subsections.

Moreover, for demonstrating the generality of DSSL, we also apply DSSL to two more SSL frameworks, MoCo v2 [4] and Barlow Twins [31], by following same settings and implementation details on SimCLR w/ DSSL. DSSL consistently benefit these two frameworks.

Our DSSL prevents closing the similarity among heavily augmented views since the commonality among these views is usually scarce, and forcing assimilating them would lead to the model collapsing. As a result, DSSL can make these views no longer risky for negative pairs free self-supervised learning methods and make them play a bigger role in contrastive learning methods which relying on negative pairs.

Linear evaluation on CIFAR-10. Similar to the implementation in the ImageNet dataset, we use the unlabeled training images in CIFAR-10 for self-supervised learning on the ResNet-18 backbone. We follow the standard settings used in the CIFAR experiments of SimSiam [5] that SGD with a learning rate 0.03, cosine lr decay schedule for 800 epochs, image size of 32×32, and a batch size of 512. We train the ResNet-18 feature encoder on unlabeled CIFAR-10 images and then freeze the backbone to train a linear tasks-specific head on CIFAR-10 with annotations.

Tab. 2 reports the top-1 accuracy of linear classifier trained on CIFAR-10. Our reproductions of SimSiam (92.1)

Method	CIFAR-10			ImageNet				
	SimCLR	BYOL	SimSiam	SimCLR	BYOL	SimSiam	MoCo v2	Barlow Twins
repro.	91.5	92.2	92.1	59.5	67.8	67.4	64.4	62.4
w/ $\widehat{\mathcal{T}}$	92.2	94.4	*collapse*	60.0	**68.3**	*collapse*	64.2	62.0
w/ DSSL	**93.2**	**94.7**	**94.5**	**60.3**	**68.3**	**68.6**	**65.3**	**62.6**

Table 2. Comparisons on linear evaluation accuracies (%). *repro*: our reproduction of each method. *collapse*: model collapsed during training. *w/* $\widehat{\mathcal{T}}$: training views are jointly augmented from standard and heavy augmentations. Heavy augmentations perform unstably even model collapsing, while DSSL *consistently* benefits from $\widehat{\mathcal{T}}$.

Method	CIFAR-10		ImageNet
BYOL (repro.)	92.2		67.8
$+\widehat{\mathcal{T}}$:	*RA*(4,10)	*RA*(8,16)	*RA*(8,16)
BYOL w/ $\widehat{\mathcal{T}}$	84.1	*collapse*	60.6
BYOL w/ DSSL	**94.4**	**94.0**	**67.8**

Table 3. Linear evaluation accuracies (%) of BYOL by applying more heavy augmentations ($+\widehat{\mathcal{T}}$).

Method	SimCLR	BYOL	SimSiam
w/ $\widehat{\mathcal{T}}$	55.7	67.1	*collapse*
w/ DSSL	**60.8**	**69.3**	**67.7**

Table 4. Results on ImageNet linear evaluation when set $\widehat{\mathcal{T}}$ as the combination of three heavy augmentations {*RA*, *UA*, *Jigsaw*}.

pre-train	COCO detection			COCO segmentation		
	AP_{50}	AP	AP_{75}	AP^m_{50}	AP^m	AP^m_{75}
ImageNet sup.	59.2	37.7	40.9	55.8	33.9	35.8
BYOL (repro.)	56.9	36.9	40.2	54.1	34.0	36.5
BYOL w/ DSSL	**57.8**	**37.6**	**40.8**	**55.0**	**34.6**	**37.0**
SimSiam†	57.5	37.9	40.9	54.2	33.2	35.2
SimSiam (repro.)	58.0	37.7	40.9	55.0	34.7	37.1
SimSiam w/ DSSL	**58.2**	**38.1**	**41.5**	**55.5**	**35.1**	**37.6**

Table 5. Results of object detection and instance segmentation fine-tuned on COCO. We adopt Mask R-CNN R50-FPN with a 1x schedule and report the bounding box AP and mask AP on COCO 2017 val. Our reproduction and DSSL versions are based on 100-epoch pre-training in ImageNet. †: the reported 200-epoch result.

and SimCLR (91.5) have higher performance than that reported in the paper of SimSiam under the same settings (91.8, 91.1). Similar to the experimental results in ImageNet, after applying DSSL, all three unsupervised learning methods **outperform the supervised ResNet-18**, whose top-1 accuracy is 93.02%.

Robustness analysis of momentum encoder. As shown in Tab. 2, the momentum encoder based method BYOL represents better robustness to the heavy augmentations than SimSiam. To study the limitation of negative sample free self-supervised methods and look into the boundary of momentum updating for the encoder, we further strengthen the magnitudes of heavy augmentations for BYOL. In particular, *RA* with more augmentation policies and higher distortion magnitudes are added. The results presented in Tab. 3 show that, although momentum encoder can partly ease heavily augmented views' side effects, such improvements are unreliability. Increasing distortion magnitudes of the augmented views would result in significantly performance dropping or even model collapse.

Robustness analysis of DSSL. According to above experimental results, DSSL steady benefit various SSL frameworks under the augmentations settings proposed in InfoMin [23]. Further, we introduce more heavy augmentation policies to study the robustness of DSSL. As shown in Tab. 4, the combination of three heavy augmentations, including *Jigsaw*, *RandAugment* and *UniformAugment* [19], are equipped, and the results support our hypothesis: the heavy augmentations play as a "double-edged sword"; Fairly contrasting all views would mislead the representation learning; DSSL is a more fundamental schema for handling heavy transformations during instance-wise self-supervised learning.

More analysis of DSSL robustness for the distortion magnitudes of $\widehat{\mathcal{T}}$ can be found in the Appendix.

Transfer to other vision tasks. We evaluate the representations benefited from DSSL on different tasks relevant to computer vision practitioners, including COCO [18] object detection and instance segmentation. Unlike linear evaluation, we fine-tune the 100-epoch pre-trained BYOL and SimSiam models end-to-end in the COCO dataset. We apply the public codebase Detectron [11] to implement Mask R-CNN [14] detector and evaluate the COCO 2017 val. Tab. 5 shows that DSSL can consistently improve the standard self-supervised learning method on downstream object detection and instance segmentation tasks.

4.3. Justification and ablation study

To understand how the asymmetric loss and partially-ordered views impact the representations learning, we design experiments with various settings of instance-wise self-supervised learning on the CIFAR-10 dataset. We apply the

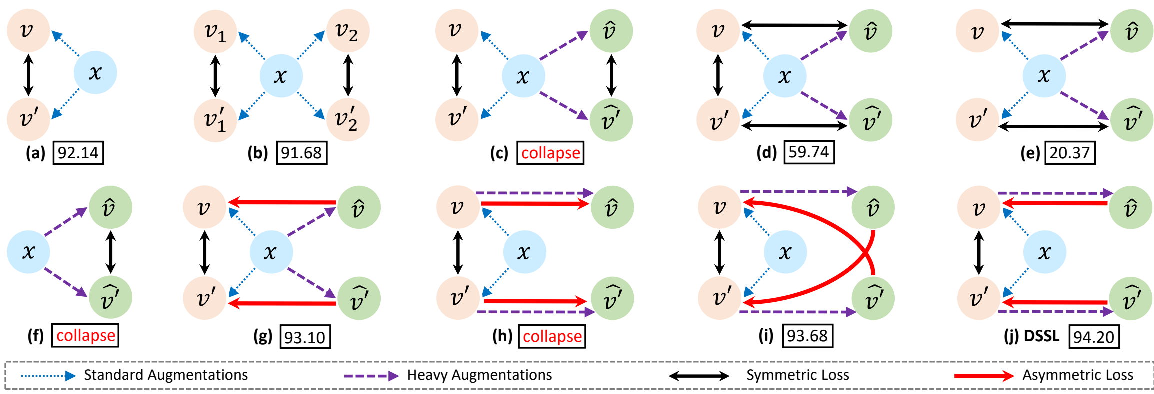

Figure 4. Comparisons of different views construction and instance-wise self-supervised learning mechanisms. Top-1 accuracies (%) of CIFAR-10 linear classifiers trained on the freeze representations are listed in the box below.

random resize-crop with color distortion as the standard image transformation, and apply *RA*(2,5) as heavy augmentation. Fig. 4 shows the *augmentation, loss function design* v.s. *linear evaluation accuracy*.

Obviously, directly making all views to be similar without distinguishing standard and heavily augmented views hurts the performances (**c**:*collapse*, **d**:59.7↓, **e**:20.4↓). According to the hypothesis proposed in InfoMin [23], it is the information about the task-relevant variable that is discarded by the hard view, degrading performance. Further, we study the impact of each components in Eq. (6) and find below four phenomenons effectively illustrate the rationality of our proposed DSSL (Eq. 5).

(i) Introducing symmetric loss $\sum_{\mathbf{V}_{\widehat{\mathcal{T}}}} \mathcal{L}_S$ among heavily augmented views results in collapsing during training (**c**, **f**).

(ii) *Fairly contrasting between the standard view and heavily augmented view has negative effects*. After comparing the results of (**h**:*collapse*) and (**j**:94.2↑), the misleading impact of $\sum_{\mathbf{V}_{\widehat{\mathcal{T}} \leftarrow \mathcal{T}}} \mathcal{L}_A$ can be observed obviously.

(iii) *Equipping directional feature prediction between heavily augmented view and standard view results in stable performance improvement* (**d**:59.7 vs. **g**:93.1).

(iv) *Partially-ordered views construction mechanism results in better performance*, as we compared among (**g**:93.1, **i**:93.7, **j**:94.2:↑). It mainly due to that the mutual information between v and its derived harder view $\widehat{v}$ is guaranteed up to a point. Such a mechanism can prevent the issue of unexpected missing information.

For more critical analysis on the influence of $\sum_{\mathbf{V}_{\widehat{\mathcal{T}} \leftarrow \mathcal{T}}} \mathcal{L}_A$, we trained a SimSiam model according to Eq. (6) by setting $\lambda + \delta = 1, \alpha = 1, \beta = 0$. As shown in Fig. 5, the negative impact of mapping heavily augmented view's feature to standard view only emerges after the value of δ raised to a threshold. At the early stage, the linear evaluation accuracy remains stable. This phenomenon further reveals mapping standard view's feature to heavily augmented view would be high-risk and low-return.

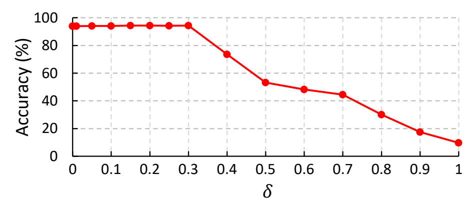

Figure 5. CIFAR-10 linear evaluation accuracy of SimSiam optimized according to Eq. (6) by setting $\alpha = 1, \beta = 0$, and $\lambda = 1 - \delta$ across varying δ. *RA* is applied to construct hard views.

Moreover, Eq. (6) is the theoretical formulation of the standard instance-wise self-supervised learning method while fairly treating all views. Our DSSL only activates two training view pairs (comparing Eq. (6) and Eq. (5)). Thus DSSL's computation complexity can be thought to be only half of the previous methods when considering all possible view pairs during training.

5. Conclusion and Discussion

We propose a directional self-supervised learning (DSSL) framework for unsupervised visual representation learning. Compared to standard self-supervised learning methods, our proposed framework benefits from more heavy image transformations and results in stable performance improvement on various vision tasks. Moreover, DSSL is easy to implement and compatible with most of the typical instance-wise self-supervised learning methods. The core concepts of DSSL can further guide the loss design of self-supervised learning. According to our formulation in Eq. (6), the soft weighted version of DSSL respecting to view characteristics is also worthy of further exploration.

Acknowledgments This work was supported by the National Key R&D Program of China under Grant No.2020AAA0108600.

References

[1] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. *Advances in Neural Information Processing Systems*, 33:12638–12648, 2020. 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1, 3, 4

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3, 4, 5, 6

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 1, 2, 3, 4, 5, 6

[6] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019. 3, 5

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 3

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 1, 3, 5

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[10] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 3

[11] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 7

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1, 3, 4, 5, 6

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[15] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *International Conference on Machine Learning*, pages 2849–2858. PMLR, 2019. 3

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[17] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 1, 2, 4

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[19] Tom Ching LingChen, Ava Khonsari, Amirreza Lashkari, Mina Rafi Nazari, Jaspreet Singh Sambee, and Mario A Nascimento. Uniformaugment: A search-free probabilistic data augmentation approach. *arXiv preprint arXiv:2003.14348*, 2020. 1, 7

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 5

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 3

[23] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. 1, 2, 3, 5, 7, 8

[24] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019. 3

[25] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *arXiv preprint arXiv:2104.07713*, 2021. 3

[26] Yu Wang, Jingyang Lin, Jingjing Zou, Yingwei Pan, Ting Yao, and Tao Mei. Improving self-supervised learning with automated unsupervised outlier arbitration. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[27] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 1, 3

[28] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, László A Jeni, and Fernando De la Torre. 3d human shape and

pose from a single low-resolution image with self-supervised learning. In *European Conference on Computer Vision*, pages 284–300. Springer, 2020. 3

[29] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019. 1, 3

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 3

[31] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 6

[32] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 3